Irreflexive and Hierarchical Relations as Translations

Antoine Bordes ANTOINE.BORDES@UTC.FR
Nicolas Usunier NICOLAS.USUNIER@UTC.FR
Alberto García-Durán ALBERTO.GARCIA-DURAN@UTC.FR
Heudiasyc UMR CNRS 7253, Université de Technologie de Compiègne, Compiègne, France

Jason Weston JWESTON@GOOGLE.COM
Oksana Yakhnenko OKSANA@GOOGLE.COM
Google, 111 8th avenue, New York, NY, USA

Abstract

We consider the problem of embedding entities and relations of knowledge bases in low-dimensional vector spaces. Unlike most existing approaches, which are primarily efficient for modeling equivalence relations, our approach is designed to explicitly model irreflexive relations, such as hierarchies, by interpreting them as translations operating on the low-dimensional embeddings of the entities. Preliminary experiments show that, despite its simplicity and a smaller number of parameters than previous approaches, our approach achieves state-of-the-art performance according to standard evaluation protocols on data from WordNet and Freebase.

1. Intoduction

Multi-relational data, which refers to directed graphs whose nodes correspond to *entities* and *edges* represent relations that link these entities, plays a pivotal role in many areas such as recommender systems, the Semantic Web, or computational biology. Relations are modeled as triplets of the form (*head*, *label*, *tail*), where *label* indicates the type of link between the entities *head* and *tail*. Relations are thus of several types and can exhibit various properties (symmetry, transitivity, irreflexivity, etc.). Such graphs are popular tools for encoding data via knowledge bases (KBs), semantic networks or any kind of database following the Resource Description Framework format. Hence, they are widely used in the Semantic Web (e.g. Freebase[1] or Google Knowledge Graph but also for knowledge management in bioinformatics (e.g. GeneOntology[2]) or natural language processing (e.g. WordNet[3]).

Despite their appealing ability for representing complex data, multi-relational databases remain complicated to manipulate because of the heterogeneity of the relations (frequencies, connectivity), their inherent noise (collaborative or semi-automatic creation) and their very large dimension (up to millions of entities and thousands of relation types).

In this paper, we introduce a distributed model, which learns to embed such data in a vector space, where entities are modeled as low-dimensional embeddings. Many existing approaches (e.g. from Sutskever et al. (2009); Nickel et al. (2011)) interpret relations as linear transformations of these embeddings: when (h, ℓ, t) holds, then the embeddings of *head* h and *tail* t should be close (in the embedding space) after transformation by a linear operator that depends on the *label* ℓ. With such an interpretation, the model implies that the relation is reflexive since the embedding of h will always be its nearest neighbor, and because of the triangle inequality, the model will, to some extent, imply some form of transitivity of the relation.

While this interpretation is fine for equivalence relations (such as WordNet's `_similar_to`), it is inadequate for irreflexive relations that represent hierarchies, such as WordNet's `_hypernym` or Freebase's type hierarchy. Indeed, taking the simplest example of entities organized in a tree with two relations, "sibling" and "parent", the embeddings of siblings should be close to each other (since it essentially is an equivalence relation), but enforcing the constraint that parent nodes should be close to their child nodes will lead the embedding of the whole tree to collapse to a small region of the space where the siblings and parent of a given node are impossible to distinguish.

[1] freebase.com
[2] geneontology.org
[3] wordnet.princeton.edu.

Since hierarchical and irreflexive relations are extremely common in KBs, we propose a simple model to efficiently represent them, by interpreting *relations as translations in the embedding space*: if (h, ℓ, t) holds, then the embedding of t should be close to the embedding of h plus some vector that depends on ℓ. This approach is motivated by the natural representation of trees (i.e. embeddings of the nodes in dimension 2): while siblings are close to each other and nodes at a given height are organized on the x-axis, the parent-child relation corresponds to a translation on the y-axis. Another, secondary, motivation comes from the recent work of Mikolov et al. (2013), in which the authors learn word embeddings from free text, and some one-to-one relations between entities of different types, such as the relation "capital of" between countries and cities, are (coincidentally rather than willingly) represented by the model as translations in the embedding space. Our approach may then be used in the context of learning word embeddings in the future to reinforce this kind a structure of the embedding space.

Apart from the main line of algorithms to learn embeddings of KBs, a number of recent approaches deal with the asymmetry of the relations at the expense of an explosion of model parameters. We present an empirical evaluation on data dumps of WordNet and Freebase, in which our model achieves strong results compared to such algorithms, with much fewer parameters and even lower dimensional embeddings.

In the remainder of the paper, we describe some of the related work in Section 2. We then describe our model in Section 3, and discuss its connections with related methods. We report preliminary experimental results on WordNet and Freebase in Section 4. We finally sketch some future work directions in Section 5.

2. Related work

Most previous methods designed to model relations in multi-relational data rely on latent representations or embeddings. The simplest form of latent attribute that can be associated to an entity is a latent class. Several clustering approaches have been proposed. Kemp et al. (2006) considered a non-parametric Bayesian extension of the *stochastic block-model* allowing to automatically infer the number of latent clusters; Kok & Domingos (2007) introduced clustering in Markov-Logic networks; Sutskever et al. (2009) used a non-parametric Bayesian clustering of entities embedding in a *collective matrix factorization* formulation. All these models cluster not only entities but relation labels as well.

These methods can provide interpretations and analysis of the data but are slow and do not scale to large databases, due to the high cost of inference. In terms of scalability, models based on tensor factorization (like those from (Harshman & Lundy, 1994) or (Nickel et al., 2011)) have shown to be efficient. However, they have been outperformed by energy-based models (Bordes et al., 2011; Jenatton et al., 2012; Bordes et al., 2013; Chen et al., 2013). These methods represent entities as low-dimensional embeddings and relations as linear or bilinear operators on them and are trained via an online process, which allows them to scale well to large numbers of entities and relation types. In Section 4, we compare our new approach to SE (Bordes et al., 2011) and SME (Bordes et al., 2013).

3. Translation-based model

We now describe our model and discuss its relationship to existing approaches.

3.1. Our model

Given a training set S of labeled arcs (h, ℓ, t), our goal is to learn vector embeddings for all values of h, ℓ and t. We assume all nodes and labels appear at least once in the training set. The embeddings take values in $\mathbb{R}^k$ (k is a model hyperparameter) and are denoted with the same letter, in boldface characters. The basic idea behind our model is that the functional relation induced by the ℓ-labeled arcs corresponds to a translation of the embeddings, i.e. we want that $\boldsymbol{h} + \boldsymbol{\ell} \approx \boldsymbol{t}$ when (h, ℓ, t) holds, while $\boldsymbol{h} + \boldsymbol{\ell}$ should be far away from $\boldsymbol{t}$ otherwise.

To learn such embeddings, we minimize the following margin-based ranking criterion over the training set:

$$\sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+ \quad (1)$$

where $[x]_+$ denotes the positive part of x, $\gamma > 0$ is a margin hyperparameter, $d(\boldsymbol{x}, \boldsymbol{y})$ is some dissimilarity function on $\mathbb{R}^k$, e.g. the euclidian distance or the squared euclidian distance, and

$$S'_{(h,\ell,t)} = \left\{(t', \ell, t) | h' \in N\right\} \cup \left\{(h, \ell, t') | t' \in N\right\}. \quad (2)$$

The set of "negative" examples we sample according to Equation 2 is basically the training ("positive") triple with either the head or tail replaced by a random entity (but not both at the same time). The loss function (1) favors low values of dissimilarity between head+label and tail for positive triplets, and large values for negative triplets, and is thus a natural implementation of the intended criterion.

The minimization is carried out by stochastic gradient descent, over the possible $\boldsymbol{h}, \boldsymbol{\ell}$ and $\boldsymbol{t}$, with the additional constraints that the L_2-norm of the embeddings of the entities is 1 (no regularization or norm constraints are given to the label embeddings $\boldsymbol{\ell}$).

3.2. Relationship to previous approaches

Section 2 described a large body of work on embedding KBs. We detail here the relationships between our model and those of Bordes et al. (2011) (Structured Embeddings or SE) and Chen et al. (2013).

SE (Bordes et al., 2011) embeds nodes into $\mathbb{R}^k$, and labels into two matrices $\boldsymbol{L}_1 \in \mathbb{R}^{k\times k}$ and $\boldsymbol{L}_2 \in \mathbb{R}^{k\times k}$ such that $d(\boldsymbol{L}_1\boldsymbol{h}, \boldsymbol{L}_2\boldsymbol{t})$ is small for positive triplets (h, ℓ, t) (and large otherwise). The basic idea is that when two nodes belong to the same edge, their embeddings should be close to each other in some subspace that depends on the label. This basic idea would imply $\boldsymbol{L}_1 = \boldsymbol{L}_2$, and using two different projection matrices for the head and for the tail is intended to account for the possible asymmetry of relation ℓ. When the dissimilarity function takes the form of $d(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x}-\boldsymbol{y})$ for some $g : \mathbb{R}^k \to \mathbb{R}$ (e.g. g is a norm), then the model of SE with an embedding of size $k+1$ is strictly more expressive than our model with an embedding of size k, since linear operators in dimension $k+1$ can reproduce affine transformations in a subspace of dimension k (by constraining the k+1st dimension of all node embeddings to be equal to 1). SE, with $\boldsymbol{L}_2$ as the identity matrix and $\boldsymbol{L}_1$ taken so as to reproduce a translation is then equivalent to our model. Despite the lower expressiveness of our model, we still reach better performance than this model in our experiments because (1) our model is a more direct way to represent the true properties of the relations, and (2) regularization, and more generally any form of capacity control, is difficult in embedding models ; greater expressiveness may then be more synonymous to overfitting than to better performance.

Another related model is the Neural Tensor Model of Chen et al. (2013). A special case of that model (which would actually boil down to a "Neural Matrix Model") corresponds to learn scores $s(h, \ell, t)$ (higher scores for positive triplets) of the form:

$$s(h, \ell, t) = \boldsymbol{h}^T \boldsymbol{L}\boldsymbol{t} + \boldsymbol{\ell}_1^T \boldsymbol{h} + \boldsymbol{\ell}_2^T \boldsymbol{t} \qquad (3)$$

where $\boldsymbol{L} \in \mathbb{R}^{k\times k}$, $\boldsymbol{L}_1 \in \mathbb{R}^k$ and $\boldsymbol{L}_2 \in \mathbb{R}^k$, all of them depending on ℓ.

If we consider our model with the squared distance as dissimilarity function, we have:

$$d(\boldsymbol{h}+\boldsymbol{\ell}, \boldsymbol{t}) = \|\boldsymbol{h}\|^2 + \|\boldsymbol{\ell}\|^2 + \|\boldsymbol{t}\|^2 - 2\big(\boldsymbol{h}^T\boldsymbol{t} + \boldsymbol{\ell}^T(\boldsymbol{t}-\boldsymbol{h})\big) .$$

Table 1. Statistics of the data sets used in this paper.

Data set	WordNet	Freebase
Entities	40,943	14,951
Rel. types	18	1,345
Train. ex.	141,442	483,142
Valid ex.	5,000	50,000
Test ex.	5,000	59,071

Considering our norm constraints ($\|\boldsymbol{h}\|^2 = \|\boldsymbol{t}\|^2 = 1$) and the ranking criterion (1), in which $\|\boldsymbol{\ell}\|^2$ does not play any role in comparing positive and negatives triplets, our model thus corresponds to the scoring triplets according to $\boldsymbol{h}^T\boldsymbol{t} + \boldsymbol{\ell}^T(\boldsymbol{t} - \boldsymbol{h})$, and thus corresponds to Chen et al. (2013)'s model (Equation (3)) where $\boldsymbol{L}$ is the identity matrix, and $\boldsymbol{\ell} = \boldsymbol{\ell}_1 = -\boldsymbol{\ell}_2$. We could not run experiments with that model, but once again our model has much fewer parameters: this should ease the training and prevent overfitting, and hence compensate for a lower expressiveness.

4. Experiments

Our approach is evaluated against the methods SE and SME (Semantic Matching Energy) from (Bordes et al., 2011; 2013) on two data sets and using the same ranking setting for evaluation.

We measure the mean and median predicted ranks and the top-10, computed with the following procedure. For each test triplet, the head is removed and replaced by each of the entities of the dictionary in turn. Energies (i.e. dissimilarities) of those corrupted triplets are computed by the model and sorted by ascending order and the rank of the correct entity is stored. This whole procedure is also repeated when removing the tail instead or the head. We report the mean and median of those predicted ranks and the top-10, which is the proportion of correct entities in the top 10 ranks.

4.1. Data

We used data from two KBs; their statistics are given in Table 1.

WordNet This knowledge base is designed to produce an intuitively usable dictionary and thesaurus, and support automatic text analysis. Its entities (termed *synsets*) correspond to word senses, and relation types define lexical relations between them. We considered the data version used in (Bordes et al., 2013). Examples of triplets are (*_score_NN_1*, *_hypernym*, *_evaluation_NN_1*) or

Table 2. Some example predictions on the Freebase test set using our approach. **Bold** indicates the test triple's true tail and *italics* other true tails present in the training set. Actual Freebase identifiers have been replaced by readable strings.

Input (Head and Label)	Predicted Tails
J. K. Rowling `influenced by`	*G. K. Chesterton*, J. R. R. Tolkien, *C. S. Lewis*, **Lloyd Alexander**, Terry Pratchett, Roald Dahl, Jorge Luis Borges, *Stephen King*, Ian Fleming
Anthony LaPaglia `performed in`	*Lantana*, *Summer of Sam*, *Happy Feet*, *The House of Mirth*, Unfaithful, **Legend of the Guardians**, Naked Lunch, X-Men, The Namesake
Camden County `adjoins`	**Burlington County**, *Atlantic County*, *Gloucester County*, Union County, Essex County, New Jersey, Passaic County, Ocean County, Bucks County
The 40-Year-Old Virgin `nominated for`	*MTV Movie Award for Best Comedic Performance*, *BFCA Critics' Choice Award for Best Comedy*, *MTV Movie Award for Best On-Screen Duo*, MTV Movie Award for Best Breakthrough Performance, **MTV Movie Award for Best Movie**, MTV Movie Award for Best Kiss, D. F. Zanuck Producer of the Year Award in Theatrical Motion Pictures, Screen Actors Guild Award for Best Actor - Motion Picture
David Foster `has the genre`	*Pop music*, *Pop rock*, Adult contemporary music, Dance music, **Contemporary R&B**, Soft rock, Rhythm and blues, Easy listening
Costa Rica football team `has position`	*Forward*, *Defender*, *Midfielder*, **Goalkeepers**, Pitchers, Infielder, Outfielder, Center, Defenseman
Lil Wayne `born in`	**New Orleans**, Atlanta, Austin, St. Louis, Toronto, New York City, Wellington, Dallas, Puerto Rico
WALL-E `has the genre`	Animations, Computer Animation, *Comedy film*, *Adventure film*, *Science Fiction*, **Fantasy**, Stop motion, *Satire*, Drama
Richard Crenna `has cause of death`	*Pancreatic cancer*, **Cardiovascular disease**, Meningitis, Cancer, Prostate cancers, Stroke, Liver tumour, Brain tumor, Multiple myeloma

(*_score_NN_2*, *_has_part*, *_musical_notation_NN_1*).[4]

Freebase Freebase is a huge and growing database of general facts; there are currently around 1.2 billion triplets. To make a small data set to experiment on we selected the subset of entities that are also present in the Wikilinks database[5] and that also have at least 100 mentions in Freebase (for both entities and relations). We also removed negative relations like '!/people/person/nationality' which just reverses the head and tail compared to the relation '/people/person/nationality'. This resulted in 592,213 triplets with 14,951 entities and 1,345 relations which were randomized and split as shown in Table 1.

4.2. Implementation

We implemented our model using the SME library[6], which already proposes code for SE and SME. The dissimilarity measure d was set to the L_1 distance, mostly because it led to a faster training.

For this preliminary set of experiments, we did not perform an extensive search for hyperparameters. For experiments of our method on WordNet, we fixed the learning rate for the stochastic gradient descent to 0.01, the dimension k of the embeddings to 20 and chosen the margin γ among $\{1, 2, 10\}$ with the validation set (optimal value was 2). We report results for SE and SME extracted from (Bordes et al., 2013) where those models have been trained using a much more thorough hyperparameter search. For experiments on Freebase, we ran all experiments using the SME library with fixed values for the learning rate ($= 0.01$), k ($= 50$) and γ ($= 1$). For both datasets, the training time was limited to at most $1,000$ epochs over the training set. The best model was selected using the mean predicted rank on the validation set.

Table 3. Link prediction results on WordNet.

Method	Rank Mean	Rank Med.	Top-10
Unstructured	317	26	35.1%
SE	1,011	**3**	68.5%
SME(linear)	559	5	65.1%
SME(bilinear)	526	8	54.7%
Our Approach	**263**	**4**	**75.4%**

4.3. Results

Tables 3 and 4 displays the results on both data sets for our method, compared to SE, to two versions of SME and to Unstructured, a simple model which only uses the dot-product between $\boldsymbol{h}$ and $\boldsymbol{t}$ as dissimilarity measure for a triplet (h, ℓ, t), with no influence of ℓ. Table 2 gives examples of nearest link prediction

[4] WordNet is composed of senses, its entities are termed by the concatenation of a word, its part-of-speech tag and a digit indicating which sense it refers to i.e. *_score_NN_1* encodes the first meaning of the noun "score".

[5] `code.google.com/p/wiki-links`

[6] `https://github.com/glorotxa/SME`

Table 4. Link prediction results on Freebase.

METHOD	RANK MEAN	RANK MED.	TOP-10
Unstructured	1097	404	4.5%
SE	272	38	28.8%
SME(LINEAR)	274	34	30.7%
SME(BILINEAR)	284	35	31.3%
Our Approach	**243**	**25**	**34.9%**

results of our approach on the Freebase test set.

Our method greatly outperforms all counterparts on all metrics, with particularly good results for the top-10 metric. We believe that such remarkable performance is due to an appropriate design of the model according to the data, but also to its relative simplicity. Hence, even if the problem is non-convex, it can be optimized efficiently with stochastic gradient. We showed in Section 3.2 that SE is more expressive than our proposal. However, its complexity makes it quite hard to train as shown in the results of tables 3 and 4.

Table 2 illustrates the capabilities of our model. Given a head and a label, the top predicted tails (and the true one) are depicted. The examples come from the Freebase test set. Even if the good answer is not always top-ranked, the predictions reflect common-sense.

5. Conclusion and future work

We proposed a new approach to learn embeddings of KBs, focusing on the minimal parametrization of the model to accurately represent hierarchical and irreflexive relations. This short paper is essentially intended to be a proof-of-concept that translations are adequate to model such relations in a multi-relational setting. It can be improved and better validated in several ways. For the experimental evaluation, this paper is the first one to present link prediction on this dump of Freebase. More benchmarking is needed, such as the comparison with models of Chen et al. (2013) and Jenatton et al. (2012). We also intend to consider learning translations of word embedding, either from free text as in (Mikolov et al., 2013) or from $(subject, verb, object)$ triplets as in (Bordes et al., 2011).

Finally, regarding modeling relations, equivalence relations in our approach are represented by a $\mathbf{0}$ translation vector, and thus enforces all members of an equivalence class to be close to each other in the embedding space (whatever the relation). Some additional degrees of freedom may be given by adding a projection matrix to each relation, so that equivalence relations only enforce entities to be close to each other in some subspace of the embedding space. However, this would increase the number of parameters, and we believe that regularization and optimization techniques should be further studied to achieve optimal performance.

Acknowledgments

We thank Thomas Strohmann and Kevin Murphy for useful discussions. This work was supported by the French ANR (EVEREST-12-JS02-005-01).

References

Bordes, A., Weston, J., Collobert, R., and Bengio, Y. Learning structured embeddings of knowledge bases. In *Proc. of the 25th Conf. on Artif. Intel. (AAAI)*, 2011.

Bordes, Antoine, Glorot, Xavier, Weston, Jason, and Bengio, Yoshua. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 2013. in press.

Chen, Danqi, Socher, Richard, Manning, Christopher D, and Ng, Andrew Y. Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. *arXiv preprint arXiv:1301.3618*, 2013.

Harshman, Richard A. and Lundy, Margaret E. Parafac: parallel factor analysis. *Comput. Stat. Data Anal.*, 18 (1):39–72, August 1994.

Jenatton, Rodolphe, Le Roux, Nicolas, Bordes, Antoine, Obozinski, Guillaume, et al. A latent factor model for highly multi-relational data. In *NIPS 25*, 2012.

Kemp, Charles, Tenenbaum, Joshua B., Griffiths, Thomas L., Yamada, Takeshi, and Ueda, Naonori. Learning systems of concepts with an infinite relational model. In *Proc. of the 21st national conf. on Artif. Intel. (AAAI)*, pp. 381–388, 2006.

Kok, Stanley and Domingos, Pedro. Statistical predicate invention. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pp. 433–440, 2007.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Nickel, Maximilian, Tresp, Volker, and Kriegel, Hans-Peter. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 809–816, 2011.

Sutskever, Ilya, Salakhutdinov, Ruslan, and Tenenbaum, Josh. Modelling relational data using bayesian clustered tensor factorization. In *Adv. in Neur. Inf. Proc. Syst. 22*, 2009.